Analysis of Excitation Source Information in Emotional Speech

S. R. M. Prasanna and D. Govind

Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, {prasanna, dgovind}@iitg.ernet.in

Abstract

The objective of this work is to analyze the effect of emotions on the excitation source of speech production. The neutral, angry, happy, boredom and fear emotions are considered for the study. Initially the electroglogtogram (EGG) and its derivative signals are compared across different emotions. The mean, standard deviation and contour of instantaneous pitch, and strength of excitation parameters are derived by processing the derivative of the EGG and also speech using zero-frequency filtering (ZFF) approach. The comparative study of these features across different emotions reveals that the effect of emotions on the excitation source is distinct and significant. The comparative study of the parameters from the derivative of EGG and speech waveform indicate that both cases have the same trend and range, inferring any of them may be used. Use of the computed parameters are found to be effective in the prosodic modification task.

Index Terms: source, emotion, pitch, strength.

1. Introduction

Emotion is the hallmark of human speech. Both emotion and naturalness provide completeness and expressiveness to speech. The emotion part of speech plays an important role in conveying the real intent and meaning of the speaker, not readily available as part of message. In human-human communication, such intent and meaning are extensively used as guiding factors for deciding the future course of action. For instance, in customer care division, the angry speech by a customer mostly indicate the dissatisfaction and hence the immediate need in the change of action. Like this the use of emotional information in human-human communication is extensive and effortless. However, the human-machine communication falls behind in this aspect significantly. There are several attempts to incorporate the same [1, 2]. However, most of the results achieved seem to be not satisfactory. Several factors may be attributed, including lack of understanding about emotions, their effect on speech production and perception systems [5]. From the engineering perspective at least what can be tried is a better understanding of the emotions by using speech signal as the carrier, setting aside the philosophical and neural aspects of emotions. Any clear understanding about the same will lead to the development better human-machine interfaces. Further, the effect of emotions on the excitation source and vocal tract system can be studied separately. This work focuses extensively on the study of the effect of emotions on the excitation source information.

The reasons for the focus on the excitation source is motivated by the following facts: Curiosity to understand how significant and systematic the effect of emotions on the excitation source. The availability of electroglogtogram (EGG) signals under different emotions [11], and fortunately good tools for estimating excitation source related parameters [7, 8, 10]. All these together may give better understanding about the changes occurring in the excitation source. The major excitation for speech production is due to the vibration of vocal folds and results in the generation of glottal wave [8]. The glottal wave indicates the modulation contour of the air and stores all the dynamics occurring in the vibration of vocal folds during speech production. The glottal wave is the ground truth and hence the best carrier to study about the effect of emotions on the excitation source. The modulation of air and hence shape of the glottal wave may be different for each emotion and can be observed by comparing the glottal waves of different emotions.

The near-periodic nature of vocal folds vibration and the strength with which the closure of vocal folds occur is better estimated computationally using the derivative of the glottal wave [9]. The associated periodicity is characterized grossly in terms of average and standard deviation values of pitch [4] and finely in terms of instantaneous pitch [7]. The strength can be characterized in terms of the amplitude of the associated epoch at the closure [8]. All these parameters may be distinct for each emotion and hence useful for analyzing emotional information of speech. These parameters can be reliably and accurately estimated using the recently developed methods for finding epochs, instantaneous pitch and strength of excitation, all based on the concept of zero frequency filtering (ZFF) [9].

The above study helps in analyzing the excitation source information in the most possible accurate way. However, in most practical speech processing tasks, the glottal wave or its derivative are not available. What is available is the speech signal and at most the Linear Prediction (LP) residual that can be used as the substitute for the glottal wave and its derivative. The next question therefore is how well the above mentioned parameters estimated from the speech and LP residual match with those using the glottal wave and it’s derivative? If the parameters from the two approaches match well, then large scale analysis can be done only using the LP residual. Finally how useful these parameters in developing human-machine interfaces? The basic essence can be demonstrated in the prosodic modification task. All these issues are studied in this paper.

The rest of the paper is organized as follows: Section 2 provides a comparative study of different emotions in terms of glottal wave and its derivative. The various parameters estimated from the derivative of glottal wave are studied in Section 3. A comparative study with the parameters estimated from the speech waveform is made in Section 4. The usefulness of estimated parameters are illustrated in the prosody modification task in Section 5. The last section provides the summary, conclusions and scope of the present work.
2. Effect of Emotions on the Glottal wave and its derivative

The significance of excitation source in generating emotions can be studied by analyzing the Electro Glotto Gram (EGG) signals of various emotions. EGG represents the actual glottal waveform that is produced by the closing and opening of the vocal folds during the speech production. The near impulse-like excitation produced due to the closing of the vocal folds forms an important characteristic of the excitation source. The location of these excitation instances is represented by impulse-like discontinuity in the derivative of the glottal waveform. Figure 1 shows the speech waveform, LP spectrum, glottal waveform and glottal wave derivative of five different emotions, namely, neutral, angry, happy, boredom and fear, for the same speaker and sound. The nature of the speech waveform is different across different emotions. The variations in the vocal tract characteristics is represented by the LP spectrum and that of excitation characteristics by the glottal wave and its derivative. Since the speaker and sound are same, the variations across different plots are primarily due to emotions. The first level visual inspection of Figure 1 indicates that the emotions significantly affect the glottal wave and its derivative as well as the vocal tract spectrum. From the derivative of glottal wave it can be seen that the amplitude of the excitation pulses (termed as strength of excitation), their periodicity (pitch contour) are also affected by the emotions. The glottal waveforms and their derivative indicate that the boredom and neutral emotions have relatively higher excitation strength and fear emotion has the least strength of excitation compared to all other emotions. Also glottal wave derivative shows that the angry and fear have the highest pitch and lowest pitch is by the boredom and neutral.

An interesting observation is the strength of excitation of angry emotion. By perception of speech we feel its strength of excitation should be largest, but comes out to be smallest. This is true because, for producing speech in angry emotion, the pitch increases and hence pitch periodicity decreases. To cater the smaller values to periodicity, the vocal folds may not be closing with high suction and hence low strength for impulse-like excitation. This interpretation may be further appreciated by analyzing the boredom case. In case of boredom, by perception of speech we feel its strength of excitation should be smallest, but comes out to be largest. This is due to the large pitch periodicity associated with boredom and hence more time for closing and opening. This leads to closing of vocal folds with high suction and hence high strength.

As characteristics of the pitch contour and strength of excitation are the significant excitation parameters that are affected by emotions. The following section describes the methods to estimate these excitation parameters directly from the derivative of EGG and are compared with those estimated from the speech waveform.

3. Estimation of Excitation Parameters

The excitation parameters like mean, standard deviation and instantaneous values of pitch and the strength of excitation can be estimated using zero frequency filtering (ZFF) method. The algorithmic steps for estimating these parameters are given below [7, 8, 9].

3.1. Epochs using ZFF method

Epochs are the impulse-like excitations corresponding to the locations of closing instants of vocal folds during vibration [6]. These can be estimated by ZFF method as follows:

- Difference input speech signal \( x(n) = s(n) - s(n-1) \)
- Compute the output of cascade of two ideal digital resonators at 0 Hz \( y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \), where \( a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1 \)
- Remove the trend i.e., \( \tilde{y}(n) = y(n) - \bar{y}(n) \), where \( \bar{y}(n) = \frac{1}{2N+1} \sum_{n=-N}^{N} y(n) \), where \( 2N + 1 \) corresponds to the average pitch period computed over a longer segment of speech
- The trend removed signal \( y(n) \) is termed as zero frequency filtered (ZFF) signal.
- The positive zero crossings of the ZFF signal will give the locations of the epochs.
3.2. Pitch Parameters [7]

- The interval between successive epochs give instantaneous pitch period.
- The average and standard deviation of instantaneous pitch values over the entire speech signal of given emotion can then be computed.

3.3. Strength of Excitation [8]

- The strength of excitation is estimated as the slope of the ZFF signal computed around each epoch location.

4. Excitation Parameters from EGG and Speech

Figure 2 plots the EGG signals, strength of excitation and pitch contours estimated from the EGG signals using ZFF method for five different emotions, for a given speaker and sentence. The envelopes of EGG signals are different indicating different way of modulation of air flow for each emotion. The variation in the nature of strength of excitation contours across different emotions indicate the difference in the rate of closure of vocal folds. It is interesting to observe that the maximum values of strength of excitation is even more compared to that of EGG. Hence the amplitudes in the two cases may represent different aspect of source information. The pitch contours are different indicating the influence of emotion on the pitch periodicity. Figure 3 represents the same excitation parameters derived from the speech signals using ZFF method. The difference in the envelopes of speech signals and strength of excitation contours are attributed to the effect of emotion. The difference between the envelopes of strengths of excitation derived using EGG and speech indicate that the observation using speech case may not directly correlate well. However, the general trend remains same in both the cases. The similarity in the pitch contours can be observed for both the cases.

Table 1: Excitation parameters of different emotions from EGG and speech.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Mean Pitch EGG</th>
<th>Std. Dev. Pitch EGG</th>
<th>Epoch Strength EGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>281.74</td>
<td>30.73</td>
<td>0.60</td>
</tr>
<tr>
<td>Happy</td>
<td>238.42</td>
<td>47.70</td>
<td>0.54</td>
</tr>
<tr>
<td>Boredom</td>
<td>168.87</td>
<td>36.34</td>
<td>0.62</td>
</tr>
<tr>
<td>Fear</td>
<td>261.61</td>
<td>45.66</td>
<td>0.53</td>
</tr>
</tbody>
</table>

4.1. Comparison of Excitation Parameters

As per Figures 2 and 3, angry and fear emotions have highest pitch contour values, whereas boredom has the lowest pitch contour values. In the case of strength of excitation, boredom has the highest, and angry and fear have the lowest values. Tables 1 provide mean pitch, standard deviation and strength of excitation computed across five sentences and three speakers using EGG and speech signals. Even though the respective values are different in the two cases, the trend is same across all the parameters. This infers that if trend is the required information, then we can derive the same either using speech or EGG. Table 2 provides the variation of excitation parameters with respect to neutral emotion. The values in the table are computed by normalizing the difference of excitation parameters of other emotion with neutral emotion. In the normalized case also the trend almost remains same in both the cases of EGG and speech. Any of these normalized parameters can be used for further processing of speech.

5. Emotion Transformation by Prosody Modification

The estimated excitation parameters can be incorporated in prosody modification task to convert speech from the given...
emotion to the target emotion. Here prosody modification involves introducing the desired pitch contour characteristics and strength of excitation in the given emotion to obtain the target emotion. The method proposed in [8] is used for the prosody modification which uses the knowledge of accurately estimated epochs. The epochs are estimated for prosody modification by ZFF method [10]. The conversion of the given emotion to neutral is done by using the parameters estimated in Table 2 as the prosody modification factors. For instance, for converting fear speech to neutral speech, the mean, standard deviation and strength of excitation have to be scaled by the factors 0.71(1-0.29), 0.83(1-0.17) and 1.22(1+0.22), respectively. The strength of excitation is introduced by scaling the Hilbert envelope of the LP residual [4] and accordingly modifying the residual prior to prosody modification. The modified strength of excitation and pitch contours for angry to neutral emotion conversion are shown in Figure 4. The mean and standard deviations of the pitch values have been scaled and now are close to the neutral case. Similarly, the strength of excitation is also scaled and resampled so that they are relatively more close to the neutral. The informal listening to the synthesized speech gave a feeling that it is sounding more as neutral than the given original emotion.

6. Summary and Future work

The objective of the present work is to analyze, estimate and incorporate the excitation parameters to obtain the emotional speech. Availability of EGG for each emotion is used for analysing the emotion dependent excitation parameters. The statistical characteristics of the instantaneous pitch contour and excitation strength contour is identified as the emotion dependent excitation parameters. Then these excitation parameters, that are estimated from the speech waveform using ZFF method are compared to above estimated parameters from the EGG. The computed excitation parameters over a data set of 3 speakers and 5 text indicated similar trends of variations in these parameters for each emotion.

7. Acknowledgements

This work is a part of ongoing UKIERI project (2007-2011) on Study of Source Features for Speech Synthesis and Speaker Recognition between IIT Guwahati, IIIT Hyderabad and CSTR, University of Edinburgh, UK. We would also like to thank Prof. Felix for providing the EGG recordings of the German emotional database.

8. References


Table 2: The normalized excitation parameters derived from EGG and speech with respect to neutral emotion.

<table>
<thead>
<tr>
<th>Exc. Parameter</th>
<th>Neutral</th>
<th>Neutral</th>
<th>Neutral</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EGG</td>
<td>Speech</td>
<td>EGG</td>
<td>Speech</td>
</tr>
<tr>
<td>Mean Pitch</td>
<td>0.38</td>
<td>0.40</td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td>Std. Dev. Pitch</td>
<td>0.41</td>
<td>0.42</td>
<td>0.27</td>
<td>0.34</td>
</tr>
<tr>
<td>Exc. strength</td>
<td>-0.45</td>
<td>-0.32</td>
<td>-0.27</td>
<td>-0.25</td>
</tr>
</tbody>
</table>

Figure 4: Strength of excitation and pitch contour of fear emotion ((a)-(b)), Modified strength of excitation and pitch contour of fear emotion ((c)-(d)), Strength of excitation and pitch contour of neutral emotion ((e)-(f)).